1000 Genomes Project: Y Chromosome SNPs

Luke Jostins, Qasim Ayub, Yali Xue, Chris Tyler-Smith

Abstract

- Y chromosome SNPs were called from the 1000 Genomes data, and numerous filters were applied
- A total of 2870 sites were called as variable in the 77 samples, of which around 75% are novel
- 30 sites that passed all filters were re-sequenced using capillary sequencing, giving an estimated false positive rate of 3.3%
- Known HapMap variants and variants from the Y haplogroup tree were used to estimate the sensitivity. This gave 22% for singletons and doubletons and 63% for variants with a non-reference allele count of three or greater
- We use the sensitivity to estimate a polymorphism rate of 1 variant per 2350bp
- HapMap genotypes and Pilot 1/Pilot 2 concordance was used to estimate a per-genotype error rate for Q10 non-reference bases of under 1%

1 Calling Variable Sites

1.1 Reads and Coverage

A total of 188M reads mapped to the Y chromosome for the 77 Pilot 1+2 males, with a mean total depth at HapMap sites of 150X, or 1.94X per individual (1.76X for Low Coverage samples, 15X for high coverage). Coverage was estimated at known genotypeable sites in order to minimize the effect of mapping errors.

There was significant heterogeneity in coverage across the samples (Figure 1).

1.2 Calling and Filtering Sites

Variant sites were called using GLFTools v3, using $\theta = 0.001$, a heterozygosity prior penalty of 50, and an RMS mapping quality threshold of 60, generating a list of 49 290 candidate SNPs. All samples had genotypes recalled at these candidate sites, using $\theta = 1$, and various filters were applied.

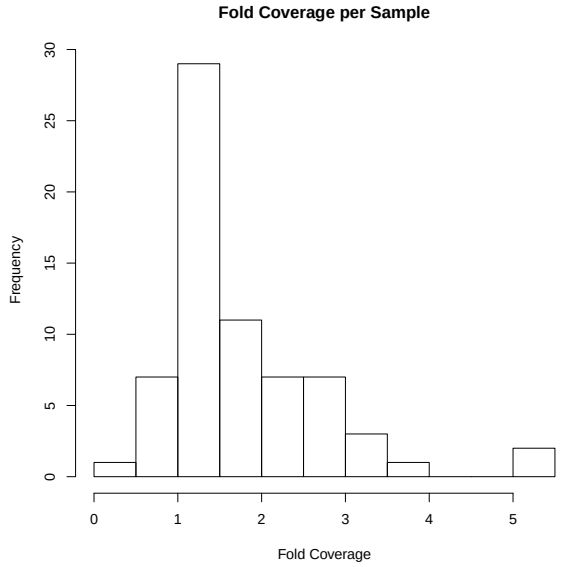


Figure 1: A histogram of Y-chromosome fold coverage per sample at 227 HapMap sites

First Stage Filters

- **Heterozygosity Filter** Remove sites with heterozygous calls of any quality (taking into account the het-prior), or more than 3 different high-certainty alleles
- **Depth Filter** Remove sites with under 1/2 or over than 3/2 times the mean depth at HapMap sites
- **Proximity Filter** Remove sites that are within 4bp of other candidate sites

Second Stage Filters

- **Haplotype Filter** Remove sites that are polymorphic in more than 1 major haplogroup
- **Quality Filter (Singletons)** Remove sites with a non-reference quality < 50
- **Quality Filter (Non-Singletons)** Remove sites with a a mean non-reference quality ¡ 30 and a total non-reference quality ¡ 100

A total of 5839 sites passed the First Stage filters, and 2870 sites passed the Second Stage filters, of which 24.1% are in dbSNP. Figure 2 shows the distribution of minor allele counts at the called sites.

A final filter was applied that flagged sites that lie outside of the approximately 12Mbp Unique Y-specific Region (UYR), though calls that failed this filter are included in the final callset. A total of 1971 SNPs fall inside this region, of which 26.0% are in dbSNP.

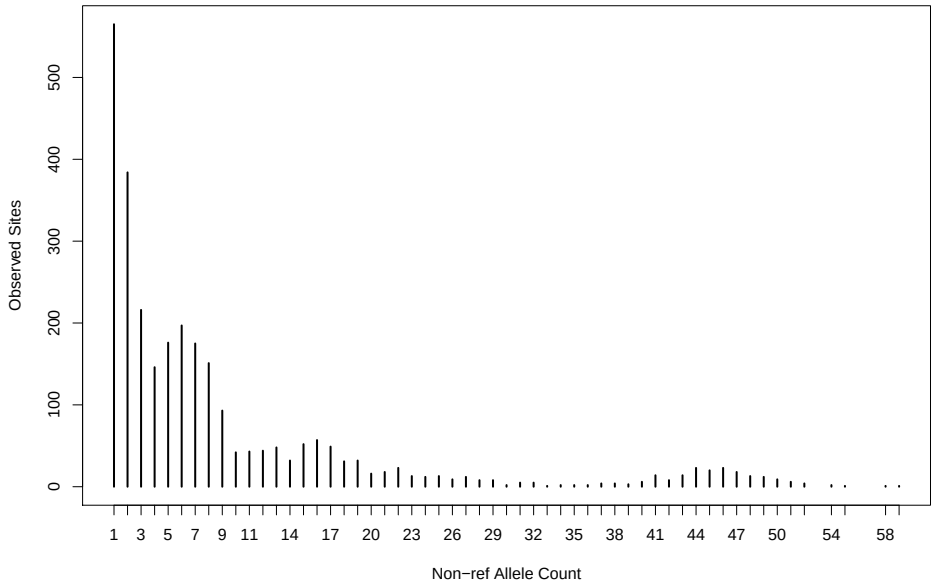


Figure 2: The distribution of non-reference allele count at the 2870 called sites

2 Validation

2.1 Capillary Sequencing of Novel Sites

200 randomly selected novel sites within the unique Y region that passed the First Stage Filters were selected for capillary resequencing. However, 4 failed primer design and 40 primer pairs were excluded as being non-specific by nucleotide BLAST (blastn). The remaining 156 were amplified in two males and a female sample and 117 that gave male specific amplimers were selected for sequencing. Of these a total of 66 sites were correctly sequenced, of which only 35 were validated.

As a result of this validation we applied the stringent Second Stage Filters. Thirty of the sequenced sites passed these filters, of which 29 validated, giving a provisional false positive rate of 3.3%, with a 95% upper confidence bound of 9.5%. A further 26 sites are currently undergoing sequencing, which will give a more accurate estimate of the False Positive rate.

2.2 Estimating Sensitivity and Polymorphism Rate

We estimated the sensitivity using 249 sites within the UYR for which could predict an expected non-reference allele count in our 77 samples; these included 173 HapMap sites and 73 sites predicted from the Y haplogroup tree. We found that 4/18 (22%) predicted singletons and doubletons were detected, and 144/229 (63%) of sites with at least three observed non-reference alleles were detected.

We can use the sensitivity, and the observed non-reference allele counts in our call set, to predict the

total number of mutations within the UYR. The predicted value is 5123, or 1 site every 2352bp. The 95% confidence interval is 1 site every 2015 to 3225bp.

2.3 Trio Concordance

Genotypes for the Pilot 1 low-coverage sequence and Pilot 2 high-coverage sequence from the Trio fathers were predicted separately; we can get a measure of the accuracy of our genotypes by examining the concordance between the two pilots. We only examined sites that were called in both Pilot 1 and Pilot 2; this consisted of 2630 sites for the CEU Trio member NA12891, and 2219 sites for the YRI Trio member NA19239.

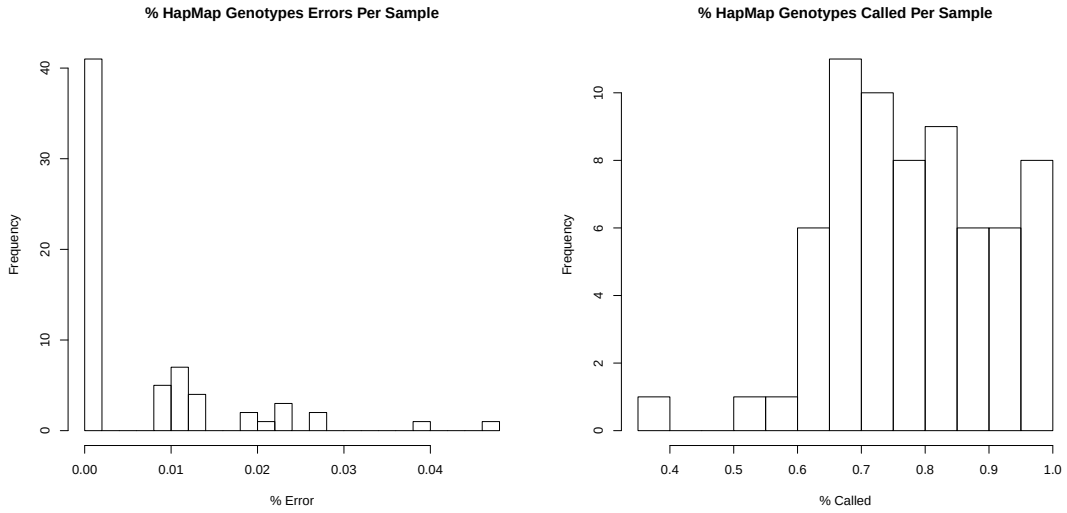
The CEU Trio-member had a 99.2% concordance, increasing to 100% for Q10; non-reference calls were slightly lower, at 97.1% and 100% respectively. The YRI Trio-member had a 98.2% overall concordance, 99.7% for Q10, and 92.4% and 98.4% respectively for non-reference calls.

The CEU accuracy was not substantially higher within the UYR. For the YRI, concordance rates were slightly higher for non-reference alleles within the UYR; 93.3% and 99.3% for all calls and Q10 calls respectively.

2.4 HapMap Concordance

We performed further validation by examining the concordance between our calls and the HapMap genotypes for the 131 sites that overlap. Of the non-missing genotypes in HapMap, 78% are genotyped in our set, and 75% have at least Q10 (these proportions are approximately the same for reference and non-reference alleles). 99.4% of calls are correct, rising to 99.9% for Q10 calls. For non-reference alleles, these proportions are 99.0% and 99.8% respectively.

As expected from the variation in coverage, there is significant heterogeneity in the proportion of genotypes calls, and the error rate, between samples (see Figure 3). Note that each sample only has, on average, 0.5 errors, so Figure 3b is subject to substantial sampling error.



(a) Call Rate

(b) Error Rate

Figure 3: Histograms of call rate and error rate per sample at HapMap sites

3 Constructing a Haplogroup Tree

A maximum likelihood haplogroup tree under a HKY model of evolution was produced using phyML, and bootstrap values were produced using 100 subsamplings. Trees were produced using both all 2870 filtered sites (Figure 4), and the 1971 UYR sites; though there was very little difference between the two trees.

The haplogroup tree classifies all the major haplogroups as monomorphic, and recovers the relationships between them, with high bootstrap confidence. It also shows evidence for a deep division between haplogroups DE and CT, previously identified only by a single marker (P143). New insights into recent human evolution can also be gained from the branch lenghts; for example, the short internal branch lengths within the haplgroup R1b relative to the other haplogroups suggest a recent expansion of this European haplogroup.

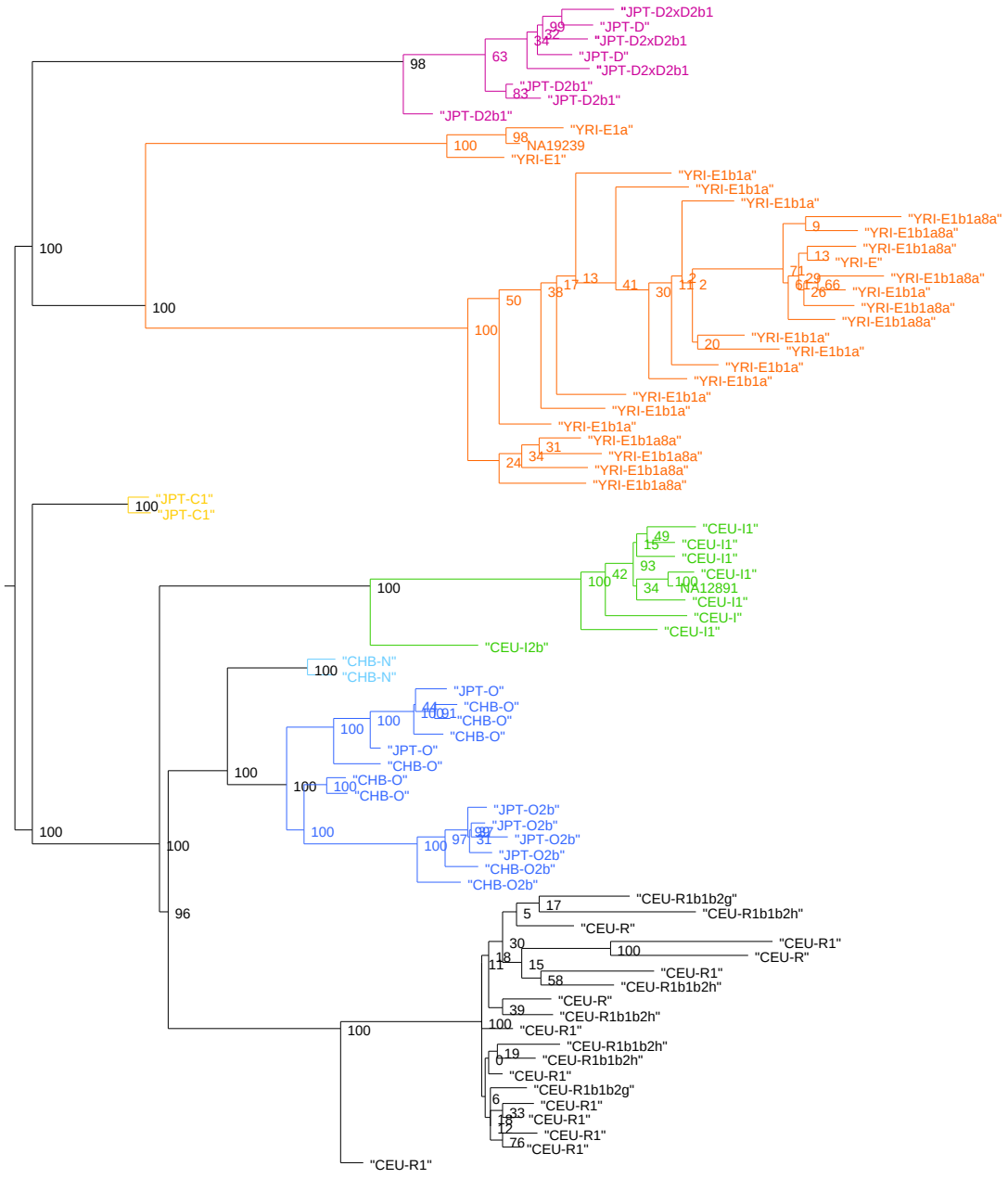


Figure 4: The Y chromosome haplogroup tree, inferred by maximum likelihood from the 2870 variable sites identified. The leaf labels contain the population and the haplogroup assignment of each sample; the latter was obtained from HapMap genotype data.